(1)

PERFORMANCE OF PERFORMANCE TESTS: COMPARISON OF PSYCHOMETRIC PROPERTIES OF 24 TESTS FROM TWO MICROCOMPUTER-BASED BATTERIES

by

Robert S. Kennedy

and

Janet J. Turnage

and

Mary K. Osteen

Essex Corporation
Orlando, Florida

March 1989

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release, distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION U.S. Army Aeromedical Research Laboratory 6b. OFFICE SYMBOL (If applicable) SGRD-UAS-NS	7a. NAME OF MONITORING ORGANIZATION U.S. Army Medical Research and Development Command
6c. ADDRESS (City, State, and ZIP Code) Fort Rucker, Alabama 36362-5292	7b. ADDRESS (City, State, and ZIP Code) Fort Detrick Frederick, MD 21701-5012
8a. NAME OF FUNDING/SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Performance of Performance Tests: Comparison of Psychometric Properties of 24 Tests From Two Microcomputer-Based Batteries (U)

12. PERSONAL AUTHOR(S)
Kennedy, Robert S., Turnage, Janet J., and Osteen, Mary K.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM ______ TO ______	1989 March	35

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	cognition, human performance, performance tests, psychomotor skills, test batteries
05	08		
13	08		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Determining whether human performance is affected by environmental stressors generally requires repeated-measures using tests which possess both stability and reliability. Cognitive and psychomotor tests are becoming popular testing devices, particularly in computer-based modes of administration. However, to date, there is no standardized battery for the study of environmental stressors which has proven itself over repeated-measures applications. In this study, three experiments were performed to evaluate the stability, reliability, and cross-test correlations of nine tests selected from the Automated Performance Test System (APTS), and 15 tests selected from the Unified Tri-Service Cognitive Performance Assessment Battery (UTC-PAB). These two batteries are in various stages of development. The APTS battery has been developed largely under NASA and NSF sponsorship. The UTC-PAB is

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Robert L. Stephens, Ph.D.	22b. TELEPHONE (Include Area Code) (205) 255-6862 22c. OFFICE SYMBOL

DD Form 1473, JUN 86 Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

19. Abstract (continued)

a DOD-mandated battery. The "best" tests are to be listed in a series of menus of factor-analytically derived batteries which will possess optimal properties of high reliability, early stability, and factorial richness at minimum costs in time. In spite of differences across studies, seven PAB tests and all APTS tests exceeded reliability and stability criteria. As a group, PAB tests took longer to stabilize but, of the tests used in this study, were more diverse factorially. Considerable overlap is present between the batteries and they have common ancestors. Suggestions are offered about how to combine the best of both menus.

Accesion For	
NTIS CRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and / or Special
A-1	

Table of contents

List of tables

Acknowledgments

The authors appreciate the role of the Biomedical Applications Research Division, especially Dr. Robert L. Stephens and Dr. Michael G. Sanders, who provided helpful guidance as Contract Officer Representatives, and LTC Gerald P. Krueger, who provided assistance with publication procedures. The authors would also like to thank Mr. Robert E. Tabler for assistance in the data collection and Mr. Martin G. Smith for programming the tests.

This material is based upon work supported by the U.S. Army Aeromedical Research Laboratory under award number DAMD 17-85-C-5095. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U.S. Army Aeromedical Research Laboratory.

This page left blank intentionally.

Introduction

Every review of human performance in every exotic environment published in the past 25 years, of which we are aware, has lamented the state of knowledge in human performance assessment: for example, underwater (Bachrach, 1975), in conditions of hypoxia (Bandaret, MacDougall, Roberts, Tappan, Jacey, & Gray, 1984), motion sickness (Hettinger, Kennedy, & McCauley, 1988), noise (Kryter, 1970), compressed gas (Bachrach, 1975), cold (Keatinge, 1969), vibration (Shoenberger, 1981), and air combat (Lane, 1986). Human performance can be affected by exposure to numerous environmental and toxic stressors in many military and nonmilitary workplaces. Environmental stressors which have been shown to alter performance include motion (Kennedy & Frank, 1986), weightlessness (Nicogossian & Parker, 1982), high altitude (Fowler, Paul, Porlier, Elcombe, & Taylor, 1985), pressure at depth (Logie & Baddeley, 1985), temperature (Ellis, 1982), and prolonged exercise combined with sleep deprivation (Rognum, Vartdal, Rodahl, Opstad, Knudson-Bass, Kindt, & Withey, 1986). Toxic effects of pharmacological agents (Mohs, Tinklenberg, Roth, & Koppell, 1970), alcohol (Kohl, Calkins, & Mandell, 1986), and mixed gases also reduce performance. Additionally, the state of the individual in terms of nutrition (Smith & Miles, 1986), chemicals (Guillion & Eckerman, 1986), physical exercise (Englund, Ryman, Naitoh, & Hodgdon, 1985), sleep loss (Woodward & Nelson, 1974), fatigue (West & Parker, 1975), and dehydration (Bandaret et al., 1984) can have an influence on performance. The effects of such stressors, however, frequently cannot be measured satisfactorily, either by self-reports or on-the-job measures, because the former can be faked and the latter are too unreliable.

A need exists for an objective, standardized cognitive and psychomotor testing tool able to detect subtle differences in the integrity of performance and welfare. A large literature survey dealing with performance effects in motor vehicles and other dynamic settings (Johnson & Kennedy, 1985) indicated that fewer than 5% of over 2000 citations reviewed addressed the effects on <u>skilled behavior</u> and the <u>basic abilities</u> underlying performance -- for example, information processing, memory, cognition, perceptual and motor skills.

Recently, considerable research effort has focused on the development of computer-based neurobehavioral and cognitive test batteries for the assessment of human performance in the presence of toxic elements and environmental stressors (e.g., Guillion & Eckerman, 1986; Baker, Letz, Fidler, Shalet, Plantamura, & Lyndon, 1985; DeRoshia (in press); Barrett, Alexander, & Forbes, 1977; Englund, Reeves, Shingledecker, Thorne, Wilson, & Hegge, 1986; Kennedy & Bittner, 1977; Shingledecker, 1984). The approach followed in the development of these batteries includes cognitive theory (Guillion & Eckerman, 1986), a desire to "standardize" (Hegge cited in Sanders, Haywood, Schroiff, & Wauschkuhn, 1986), the need to field a battery which tests the mental functions disrupted by the agents (Baker,

et al., 1985) and the opportunity to computerize paper-and-pencil tests (Barrett et al., 1977). To this list we add our own philosophy.

We believe that, instead of following cognitive theory, the technical requirements for developing a battery of tests should be based on the tenets of the classical theory of mental tests and testing (e.g., Allen & Yen, 1979; Gulliksen, 1950; Thorndike & Hagen, 1977). Test theory requires that tests meet set criteria like stability (or parallel forms) and reliability (the lack of which constitutes insensitivity). This is the approach that we have followed in the development of the Automated Performance Test System (APTS).

The APTS battery

Several years ago, under National Aeronautics and Space Administration sponsorship, a program was established to develop a battery of tests for repeated-measures application and has been described by Bittner, Carter, Kennedy, Harbeson, & Krause (1983). This work followed the earlier (1976-1981) work of Performance Evaluation Tests for Environmental Research (PETER) (Bittner, Carter, Kennedy, Harbeson, & Krause, 1983). Initially, the repeated-measures stability and reliability of paper-and-pencil tests (Carter, Kennedy, & Bittner, 1980) were compared to those of the same tests delivered via a microbased computer (Kennedy, Wilkes, Lane, & Homick, 1985; Kennedy, Wilkes, & Kuntz, 1986). Stabilities and reliabilities of microbased tests were generally high and correlated well with paper-and-pencil versions, suggesting the feasibility of computerized testing.

Over 50 reports have been written describing the development and use of the APTS for repeated-measures study of human performance, most of which are reviewed in Kennedy, Wilkes, and Baltzley (in press). The menu of tests in the APTS have the following characteristics: (1) they stabilize very rapidly (i.e., in less than three sessions); (2) they have high levels of reliability (i.e., $\underline{r}$ = > 0.70 for three minutes); (3) the tests tap a wide range of factors (i.e., three or more factors using six tests and 10 minutes' testing time); (4) they are implemented on a portable microcomputer (either a NEC PC8201A or an IBM compatible Zenith 181/183); (5) they can be used under a wide variety of field conditions (e.g., hypo- and hyperbaric chambers, ships at sea, mountain cabins, and tunnels); (6) they possess predictive validity in that different tests in the menu have been shown to be related to different global measures of intelligence (Wechsler Adult Intelligence Scale-WAIS, American College Testing-ACT, Wonderlic Personnel Test-WPT, and Armed Services Vocational Aptitude Battery-ASVAB); (7) they possess construct validity in that to some extent they have been factor analyzed and can be related to factor-based marker tests from other batteries; and (8) several of the tests have been shown to be sensitive to different environments (hypoxia, alcohol, drugs, sleep loss). Test theory has been followed as an experimental approach in APTS development. The most important practical aspects of these requirements from our point of view are discussed below.

What the tests of the APTS test, and how they fit into a conceptual model of cognition, has been left until later in development.

The UTC-PAB battery

In 1986, another test battery, the Unified Tri-Service Cognitive Performance Assessment Battery (UTC-PAB) (Englund et al., 1985) was introduced and mandated to be the primary authorized Department of Defense (DoD) test battery for assessment of cognitive performance by the Joint Working Group on Drug Dependent Degradation of Military Performance (JWGD3 MILPERF). The UTC-PAB was similar to the Performance Assessment Battery developed by the Walter Reed Army Institute of Research (WRAIR-PAB) (Thorne, Genser, Sing, & Hegge, 1985), although several tests from the Navy (Naitoh, 1982; Kennedy & Bittner, 1977) and Air Force (Shingledecker, 1984) batteries were also included. To date, the WRAIR PAB has been used in experiments studying the effects of sleep deprivation, sustained performance, jet lag, physical fatigue, hypoxia, sickle cell disorders (Thorne, 1982), and heat stress (Hamilton, Simmons, & Kimball, 1982; Mitchell, Knox, Edwards, Schrimscher, Siering, Stone, & Taylor, 1985). Recently, they too have begun to be implemented on a microcomputer (Reeves & Thorne, 1988).

The UTC-PAB, consisting of 25 cognitive and psychomotor tests that were selected from test batteries currently in existence throughout DoD research laboratories (Englund et al., 1986), was developed to provide a standardized, sensitive instrument. Although extensive background literature exists for each test of the UTC-PAB (Perez, Masline, Ramsey, & Urban, 1987), little or no information exists regarding the stability, reliability, and correlations among these tests in the newly computerized form, nor have they exhibited sensitivity in published studies to date. Answers to these questions require evaluation by repeated-measures design. Therefore, since a considerable body of empirical data were available for well over 100 tests from the PETER/APTS work, we decided to compare those tests in a core battery to an assortment of UTC-PAB tests similarly implemented. It was recognized that, since each battery was largely derived from the same corpus of information, there would be considerable overlap.

Technical requirements

The typical paradigm in environmental and behavioral toxicology research involves the use of a repeated-measures technique (ordinarily preceding, during, and subsequent to exposure to the treatment). A repeated-measures design is more efficient and economical than alternative approaches (Winer, 1971) and is ideally suited to experiments conducted with a small number of subjects. Because sample size and replications can be traded off for the same levels of statistical power (Dunlap, Jones, & Bittner 1983), it may be unethical to use more subjects than necessary in hazardous or stressful studies if such data, available on the same group, would serve the desired statistical objective.

Stability

Repeated-measures studies require stable measures if changes in the environment are to be meaningfully related to changes in performance (Jones, 1980). Of particular concern is the fact that a subject's scores may differ significantly over time. Stability of means and standard deviations are therefore important requirements, but stability of individual differences are also needed. When cross-session correlations are nonsymmetrical and exhibit superdiagonal form (Jones, 1979) it is problematic as to whether the subject or the task is changing; in either case, the performances are unstable (Alvarez & Hulin, 1972). Other test developers focus on stability of means and standard deviations and ignore symmetry of the variance/covariance matrix. We have reported instances where means and standard deviations appeared stable, but correlations were not (McCauley, Kennedy, & Bittner, 1980; Kennedy & Bittner, 1977). Jones (1980), for example, maintains that the advancement of a skill involves an acquisition phase in which persons improve at different rates, and a terminal phase in which persons reach or approximate their individual limits. Thus, when the terminal phase of performance is reached scores will cease to deviate despite additional practice. Unless tests have been practiced toward this point of differential stability, the determination of changes in scores due to practice or some other variable would be impossible. Although skills are the human behavior most often implied when acquisition is discussed, cognitive ability tests have shown the same time course changes with practice (Anderson, 1985). Our experience (Bittner, Carter, Kennedy, Harbeson, & Krause, 1983) has also shown that many of these cognitive tests may also be unstable. So far as we know, individual differences in rate of learning are not controlled nor provocatively studied in other test battery development programs, although doubtless such concerns should be considered.

Task definition

Task definition is defined as the reliability of the stabilized task (Jones, 1979, 1980), and is calculated as the average intertrial correlation between testing trials following the trial when "differential stability" occurs. When variances are constant, higher reliability (i.e., task definition) improves power in repeated-measures studies. Low reliability assures insensitivity of a test. So far as we know, no other battery places such emphasis on the requirement for high reliability.

Task ceiling

If all or several subjects obtain the maximum level of performance, then the task is said to have a ceiling (Jones, 1980). Ceilings (and floors) are undesirable because such a test limits discrimination between subjects. Often, tests will reveal their defect through a gradual reduction in between-subject variance.

Factor richness

Finally, because different agents may interact with different aspects of performance, tasks which possess the features listed above should not overlap; each test added should contribute as much new variance as possible.

Stabilization time

When all tests are stable, the more desirable performance measures are those which stabilize more rapidly following brief periods of practice without forfeiting metric qualities. Practical considerations for tasks under consideration for environmental research must account for the number of trials necessary to establish stability.

Reliability efficiency

Test reliability is influenced by test length (Guilford, 1954); other things being equal, tests with longer administration times and/or more items maintain a reliability advantage over those with less. Thus, test lengths must be equalized before meaningful comparisons can be made. A useful tool for making such relative judgments is called the reliability-efficiency, or standardized reliability, of the test (Kennedy, Carter, & Bittner, 1980), which is computed by correcting the reliabilities of different tests to a common test length or time by use of the Spearman-Brown prophecy formula (Guilford, 1954, p. 354). We suggest that ability tests should not be considered to be reliable unless they reach $r = 0.707$ for a three-minute session, which means that the test-retest variance accounted for equals 50%.

In summary, there are three explicit metric requirements and these follow from classical test theory (Allen & Yen, 1979). All tests included in the battery must exhibit stability and reliability and lack a task ceiling. From these follow the practical issues of economy of testing times, so that stability, reliability, and diversity of factors should be achieved in a minimum of time and tests. Adherence to these requirements is often ignored because they are difficult to accomplish. For example, repeated-measures experiments and factor analysis studies are each costly and those costs grow geometrically as experiments are conducted to satisfy both needs simultaneously. Not infrequently, lessons learned from one focus of study impact on the other. As a practical matter, repeated measures (trials) are not found in the scientific literature often; sample size is increased to address factor issues because they are theoretically more pleasing. This is a mistake, not only because of the effect on test battery development, but also because most tests do not stabilize immediately and factor structure is likely to change with practice. Yet definitive studies are almost totally absent in the literature (e.g., Fleishman & Hempel, 1955; Jones, Dunlap, Bilodeau, 1984; Alvares & Hulin, 1972).

Therefore, the ultimate goal of this study was implementation of the tests of both APTS and UTC-PAB batteries on the same computer, to compare their metric properties and practical features in order to determine comparability and equivalence and to suggest an optimal set(s) of tests for environmental testing.

Method

Subjects

In Studies 1 and 2, 25 right-handed male university students were recruited. In Study 3, 25 students were recruited without these restrictions. Nine males and 16 females applied. All participants were solicited on a voluntary basis in accordance with American Psychological Association principles for research with human subjects (American Psychological Association, 1982). Subjects were advised they would need to participate one-half hour each working day for a three-week period in Study 1, for three-fourths of an hour each working day for a three-week period in Study 2, and for one hour each working day for a two-week period for Study 3. Subjects were paid at the rate of $5.00 per hour. In each study, some subjects attrited due to schedule conflicts, so that final analyses were based on data from 23 subjects for Study 1, 21 for Study 2, and 24 for Study 3. The subjects were in good physical and mental health and varied from freshman to senior standing. Motivation for the research task appeared to remain high throughout the experimental sessions.

Materials

In Study 1, six tests were selected from the UTC-PAB (Englund et al., 1986) which were adaptable to a battery-operated portable microcomputer testing mode, demonstrated conformity to general criteria for "good" performance tests, and indicated well-differentiated factors. A full listing of selection criteria is given in Turnage, Kennedy, and Osteen (1987). For each PAB task in which there was a choice of demand conditions (low, moderate or high), the low demand condition was used. (For the sake of brevity, for the remainder of the article, the UTC-PAB will be referred to as PAB, but should not be confused with the original WRAIR-PAB). The six selected tests were: Mathematical Processing, Continuous Recall, Memory Search, Matrix Rotation, Successive Pattern Comparison, and Item Order.

These same selection criteria were used to select and assess five UTC-PAB tests in Study 2 (Grammatical Reasoning, Code Substitution, Matrix Rotation, Visual Scanning, and Four-Choice Reaction Time), and eight UTC-PAB tests in Study 3 (Manikin, Memory Search, Symbolic Reasoning, Vertical Addition, Time Wall, Simultaneous Pattern Comparison, Matrix Rotation, and Mathematical Processing). Complete descriptions of each of these 15 UTC-PAB tests is given in extensive government reports (Tabler, Turnage, & Kennedy, 1987; Turnage et al., 1987; Turnage, Kennedy, Osteen, & Tabler, 1988) as well as in Englund et al. (1986). Studies 1 and 2 were

conducted essentially in parallel and Study 3 benefitted to considerable extent from preliminary analyses of Studies 1 and 2. Therefore, this afforded the opportunity to retest several UTC-PAB tests (e.g., Matrix Rotation, Mathematical Processing, Memory Search) more than once and to determine whether changes in administrative procedures would alter results.

Six cognitive performance tests from the APTS (Bittner, Smith, Kennedy, Staley, & Harbeson, 1985; Kennedy, Wilkes, Dunlap, & Kuntz, 1987) were also included for comparison and examination in each of the three studies. The APTS tests were: Grammatical Reasoning, Four-Choice Reaction Time, Two-Choice Reaction Time, Simultaneous Pattern Comparison, Manikin, and Code Substitution. Two-Choice Reaction Time was eliminated from Study 3 because it was determined to be redundant. In addition, a series of 10-second finger tapping exercises was included in the test batteries as a check against interfering factors such as fatigue or boredom during battery administration and to test fine motor skills. A more complete description of each of these tests is also given in previous technical reports (Tabler et al., 1987; Turnage et al., 1987, 1988).

The UTC-PAB manual (Englund et al., 1986) lists 18 response measures which may be collected for most UTC-PAB tasks. All 18 measures were recorded for each PAB test for each study, and 11 comparable measures were recorded for APTS tests. However, only number correct, percent correct, and average response latency measures were analyzed in Studies 1 and 2, and only number correct and response latency measures were analyzed in Study 3. The PAB Time Wall task was evaluated by a time difference measure, which recorded the difference from a calibrated time standard. Tapping task measures were the number of alternate key presses, and Reaction Time tasks used only the response latency measure.

Feedback (knowledge of results) was provided to subjects for the PAB in both orientation and testing sessions for Studies 1 and 2, as prescribed by the instructions, and for the APTS only during practice (Englund et al., 1986). In Study 3, feedback was furnished to participants for both batteries during the orientation session but not during the ensuing sessions. Table 1 summarizes the administrative differences among the three consecutive studies.

Table 1.

Summary of administrative differences among the three studies

		Study 1		
No. Ss	Testing Period	Apparatus	Feedback	Response Measures
23 males	3 wks, 15 sessions PAB = 11 trials APTS = 8 trials	NEC PC8201A	Orientation PAB Testing PAB	Number Correct Percent Correct Response Latency
		Study 2		
21 males	3 wks, 15 sessions PAB = 15 trials APTS = 15 trials	NEC PC8201A	Orientation PAB Testing PAB	Number Correct Percent Correct Response Latency
		Study 3		
9 males	2 wks, 10 sessions	Zenith ZFL181	Orientation	Number Correct
15 females	PAB = 10 trials APTS = 10 trials	(Augmented by Smart System)	PAB, APTS Testing, none	Response Latency

Apparatus

In Studies 1 and 2, the testing was conducted using six NEC PC8201A microprocessors which are fully described in NEC Home Electronics (1983) and Essex Corporation (1987). Use of this microcomputer made possible the full automation of test presentations and the recoding, scoring, and storage of responses. In Study 3, the Zenith Data Systems ZFL-181 microprocessor was used with tests programmed in the Microsoft Quick BASIC. In a comparison study of NEC and Zenith microcomputer administrations (Wilkes, Kennedy, & Kuntz, 1987) over 20 microcomputer-based performance tests were practiced to stability. While there were significant differences in mean scores on some of the tests administered between computers, there were no marked differences in the cross-correlations of test scores. Except for the mean differences, it was tentatively concluded that the two computers would generate comparable results factorially, although it is necessary to attend to the future prospect of difficulties in this area.

Procedure

Prior to testing, subjects received a brief introduction to the purpose of the study and were advised regarding the general procedures associated with data collection. Subjects were directed to respond quickly, accurately, and to the best of their abilities and to follow specific instructions proposed by the respective test. During the one-hour orientation (training) session which preceded testing in each of the three studies, subjects practiced each test and were allowed to ask questions to resolve any difficulties. The data indicated that some subjects in Studies 1 and 2 apparently failed to understand instructions since they exhibited random responses (50% correct levels even with a proctor present). Therefore, in Study 3, during the orientation session, a "Smart" warning system was implemented. This computerized system directed the subject to "see the experimenter" if the subject failed to obtain a score greater than 60% correct or answered incorrectly on five consecutive trials for any test, effectively forcing participants to follow instructions during the one-day orientation session. The system was seldom reactivated after the first session.

In Study 1, subjects were examined over a 3-week period, using one PAB series on Days 1, 2, 3, 4, 5, 6, 7, 9, 12, 14, and 15, and two APTS series on Days 8, 10, 11, and 13. Although a "better" experimental design may have entailed total symmetrical interdigitation of all tests, practical issues of testing time and limited apparatus availability dictated the approach we settled upon. Because less was known about practice requirements for PAB in this, the first study in this series, it was decided that monitoring the learning curve for 11 sessions would be most informative. In addition, the uninterrupted administration of PAB for seven trials allowed the PAB to reach stability before introducing APTS which we believed from past research would stabilize early.

In Study 2, participants were examined over a four-week period, with half of the participants operating under an ABBA design (i.e., the APTS series followed PAB series on odd days and PAB followed the APTS on even days), and the other subjects operating under a BAAB design (i.e., the PAB series followed by the APTS series on odd days and the APTS followed by PAB on even days). Testing was administered throughout a two-week period followed by a one-week lay off, which was followed by the concluding week of testing. There was a one-week lay off between Trials 10 and 11 to determine if disruption of practice interfered with performance. Therefore, all participants performed both batteries during each of the 15 testing sessions. In Study 3, participants were examined over a 10-session period with each participant receiving first the PAB and then the APTS series of tests. Aggregate experimental time-on-task minus practice time for the batteries represented in each study ranged from 4.8 hours for Study 1, 4.3 hours for Study 2, and 5.3 hours for Study 3.

Four to six subjects were scheduled to report each hour. Experimental rooms contained separate tables and chairs to accommodate subjects.

Testing could generally be accomplished within one hour. If a subject were unable to report for testing at the scheduled time, rescheduling was accomplished using the criteria that no more than two testing trials could be accomplished on any one day and the order of testing needed to be preserved. The presentation order, practice times, individual trial times, and total battery administration times for both PAB and APTS series are reported in previous reports (Tabler et al., 1987; Turnage et al., 1987; Turnage et al., 1988). Although practice times were generally similar across test batteries and consisted of 30-second presentations of tests, excluding Tapping which was presented for 10 seconds, individual trial times varied for the two test batteries. All PAB tests were presented for 180 seconds per trial with the exception of Time Wall which received 110 seconds per trial. APTS tests were presented for 90 seconds per trial except for Grammatical Reasoning and Pattern Comparison which were presented for 105 seconds per trial.

Analyses

Group means and standard deviations for each individual test and for each response measure were examined for anomalies and for evidence of test stabilization, and associated intersession correlations were assessed for evidence of differential stability after the methods of Jones (1970, 1980). Two analysts independently selected the trial at which means, standard deviations, and intersession correlations appeared to plateau. There were few disagreements, and those which occurred were discussed to arrive at a consensus decision.

For each study, after the trial of stability was determined, the estimated stabilized reliability (i.e., the average intertrial correlation for all trials including and following the trial of stability) was calculated. APTS tests varied in length and thus reliabilities were adjusted according to the Spearman-Brown (Winer, 1971, p. 286) prediction equation to normalize them to the three-minute base, which is the standard administration time length for PAB tests.

To determine test overlap indicative of common factor structure, intercorrelations were computed to yield a matrix for all tests and test scores. In this intercorrelational analysis, only stable trials were used. In Study 1, the most stable trials which were selected to generate the intercorrelational matrix of tests and test scores were PAB trials 7 and 8, and APTS trials 5 and 6; in Study 2, trials 9 and 10 were selected, and trials 7 and 8 were chosen for Study 3. These particular trials were selected because scores as a rule did not stabilize until the second to fourth day of testing and tended to be more erratic on the last day(s) of testing (cf., Carter, Krause, & Harbeson, 1986). Thus, these selected trials occurred at approximately the middle of the stabilized sessions, and the APTS and PAB scores when correlated provide information about stabilized reliabilities and test intercorrelations. The correction-for-attenuation formula (Spearman, 1904) was applied on the

intercorrelations of stabilized trials to determine an indication of overlap among measures.

Applications of these formulae produced screening criteria for evaluation of both PAB and APTS tests using as requirements whether the test was stable, reliable, and the extent to which it measured a unique factor (i.e., was not highly correlated with other tests) after correcting for whatever unreliabilities were present.

Results

Tables 2, 3, and 4 present a comparison of the estimated trial of stability for means, standard deviations, and intertrial correlations, respectively, for the response measures which were assessed for each test across the three studies. Table 5 shows the estimated test reliabilities for each test across studies. Last, Tables 6 and 7 report the number of cross-task correlations between the PAB and APTS tests for the three studies with Spearman-Brown and attenuation formulas applied to number correct and response latency measures, respectively. The results reported in each table will be discussed in turn. Original data for these tables appear in more detail elsewhere (Tabler et al., 1987; Turnage et al., 1987; Turnage et al., 1988).

Table 2.

Comparison of estimated trial of stability
for means across three studies of PAB and APTS

	Study 1			Study 2			Study 3	
Test	NC*	PC	RL	NC*	PC	RL	NC*	RL
PAB								
Code Substitution				6	3	6		
Matrix Rotation	7	2	7	3	3	2	4	2
Recall	7	2	7					
Math Processing	7	2	7				3	3
Item Order	7	7	3					
Memory Search	4	2	4				4	2
Pattern Comp (Succ)	2	2	3					
Reasoning (Gram)				7	2	4		
Reaction (4)				5	1	2		
Symbolic Reasoning							2	2
Vertical Addition							3	3
Time Wall								2
Pattern Comp (Simult)							2	2
Manikin							3	3
Visual Scanning				3	2	3		
APTS								
Tapping (Pref)	3			2			2	
Pattern Comp (Simult)	4	1	2	3	2	3	5	2
Manikin	4	2	4	3	4	3	5	3
Reasoning (Gram)	4	1	2	3	5	5	3	5
Reaction Time (4)			4			2		1
Reaction Time (2)			1			2		
Code Substitution	4	3	4	3	2	2	4	3
Tapping (Nonpref)	3			2			3	
TF Tapping	3			2			2	

*Tapping tests measures as number of alternate hits.
Note: NC = Number Correct; PC = Percent Correct; RL = Response Latency

Table 3.

Comparison of estimated trial of stability
for standard deviations across three studies of PAB and APTS

	Study 1			Study 2			Study 3	
Test	NC*	PC	RL	NC*	PC	RL	NC*	RL
PAB								
Code Substitution				2	2	2		
Matrix Rotation	1	1	6	3	2	2	2	1
Recall	1	1	1					
Math Processing	3	1	3				3	3
Item Order	1	1	1					
Memory Search	3	1	2				1	1
Pattern Comp (Succ)	2	1	2					
Reasoning (Gram)				2	2	2		
Reaction (4)				2	2	2		
Symbolic Reasoning							1	2
Vertical Addition							2	3
Time Wall								3
Pattern Comp (Simult)							1	1
Manikin							3	2
Visual Scanning				2	2	2		
APTS								
Tapping (Pref)	3			2			2	
Pattern Comp (Simult)	1	1	1	2	2	2	2	2
Manikin	1	1	1	2	2	2	2	3
Reasoning (Gram)	--	1	1	2	2	2	3	5
Reaction Time (4)			5			2		1
Reaction Time (2)			3			2		
Code Substitution	1	1	1	2	2	2	1	1
Tapping (Nonpref)	1			2			2	
TF Tapping	1			2			2	

*Tapping tests measures as number of alternate hits.
**Did not stabilize.
Note: NC = Number Correct; PC = Percent Correct; RL = Response Latency

Table 4.

Comparison of estimated trial of stability
for intertrial correlations across Ttree studies of PAB and APTS

	Study 1			Study 2			Study 3	
Test	NC*	PC	RL	NC*	PC	RL	NC*	RL
PAB								
Code Substitution				7	5	--		
Matrix Rotation	2	2	2	2	2	5	2	2
Recall	2	--	7					
Math Processing	2	9	2				1	1
Item Order	2	7	2					
Memory Search	2	--	2				3	3
Pattern Comp (Succ)	7	--	7					
Reasoning (Gram)				7	7	7		
Reaction (4)				7	7	--		
Symbolic Reasoning							3	1
Vertical Addition							3	3
Time Wall								2
Pattern Comp (Simult)							4	5
Manikin							2	1
Visual Scanning				2	3	6		
APTS								
Tapping (Pref)	2			3			2	
Pattern Comp (Simult)	2	3	3	3	2	6	5	2
Manikin	3	3	3	4	3	3	2	3
Reasoning (Gram)	4	2	2	3	7	7	1	1
Reaction Time (4)			4			2		1
Reaction Time (2)			4			7		
Code Substitution	3	5	3	3	3	3	2	3
Tapping (Nonpref)	2			2			3	
TF Tapping	2			2			1	

*Tapping tests measures as number of alternate hits.
**Did not stabilize
Note: NC = Number Correct; PC = Percent Correct; RL = Response Latency

Stability of means

Table 2 indicates that test means generally stabilized within the testing periods for all response measures (number correct, percent correct, and response latency). Almost no test stabilized on the first trial but several appeared to stabilize by trial 2. Overall, PAB tests stabilized slightly later than APTS tests, especially in Study 1. This later stabilization of PAB tests in Study 1 is undoubtedly a function of the PAB series having been introduced and practiced prior to introduction of the APTS series on Day 8. However, APTS tests also tended to stabilize on earlier trials than PAB tests in Study 2, even though APTS tests were always shorter than PAB tests. In Study 3, PAB tests stabilized on earlier trials than APTS tests although after an equivalent amount of practice. Note that APTS tests were always shorter than PAB tests. The Tapping test means stabilized consistently early. Means for percent correct measures stabilized relatively early for most tests, while the number correct means consistently stabilized later than the response latency means for both batteries. There was the suggestion that the tests with the highest response rates (e.g., tapping) stabilized early and the ones with the lowest response rates (e.g., Grammatical Reasoning) stabilized latest on both APTS and PAB tests.

Stability of standard deviations

The trial of stability for standard deviations shown in Table 3 occurred early across studies. Some standard deviations showed lower variability on early trials and increased in later trials, a not uncommon finding (cf. Carter, Kennedy, & Bittner, 1981) and are not considered indicative of instability. Overall, there appeared to be smaller between-subject differences on PAB measures than in the APTS measures in Study 1 and 2, but these were not evident in Study 3 where subjects had received prior training with the "Smart" system.

Differential stability

Inspection of the trial-to-trial intercorrelations of a test provides information about stability as well as reliability. Differential stability, the point at which correlations stabilize, is best characterized as that point at which between-subject differences over sessions become parallel and thus provides an estimate of stable retest reliabilities. Table 4 indicates the estimated trial of differential stability (i.e., where intertrial intercorrelations plateau). Where intertrial correlations failed to stabilize within the testing period, a dash (-) appears in Table 4. Determination of the trial of stability was made objective in Study 3 by application of an analytic approach which is described in Turnage et al. (1988). In general, the estimated trial of differential stability was quite similar to associated trials of stability for means and standard deviations for each test and occurred slightly earlier for Study 3 than for previous studies. This result was probably

due to the more objective trial specification approach as well as to the operation of the "Smart" system which forced subjects to understand test procedures before being tested.

Some PAB measures did not exhibit stability, but all APTS did. Generally, differential stability took longer to attain than either stability of means or standard deviations. There was a slight advantage for number correct over latency scores and both took less practice than percent correct scores. APTS tests, which were always shorter on the average, also stabilized on an earlier trial than PAB tests. However, if number correct is considered the metric of choice, most individual tests from both batteries stabilize early. Again, it appeared that the difficult and more complex tests of APTS and PAB took longest to stabilize and the simpler tasks shorter.

Table 5 reports the estimated stable test reliabilities for PAB and APTS tests based upon the average intercorrelation for all trial comparisons including and following the trial of stability. The percent correct measure showed generally lower reliability than number correct and response latency measures across tests in Study 1, and did not stabilize at all for three PAB tests. Because reliabilities for the percent correct measure were similarly low for Study 2, a not uncommon finding (cf., Seales, Kennedy, & Bittner, 1980), the measure was dropped for purposes of stability analysis in Study 3.

There were large differences between observed reliabilities for the PAB and APTS series in Study 1, the average PAB reliability across measures being $\underline{r} = 0.52$ compared with the average APTS reliability across measures of $\underline{r} = 0.81$ (excluding Tapping). In Study 2, the average PAB reliability was $\underline{r} = 0.58$ compared with the average APTS reliability of $\underline{r} = 0.76$. Study 3 resulted in a small difference between batteries with PAB $\underline{r} = 0.79$ and APTS $\underline{r} = 0.82$. The superiority of APTS in Study 1 may be due to the fact that the APTS was not introduced until the eighth day of testing. However, when test batteries were presented in counterbalanced order in Study 2, there was still a noticeable difference between reliabilities of the two batteries. Indeed, if number correct is selected as the metric of choice all APTS reliabilities were higher than all PAB reliabilities. However, when the "Smart" system was implemented in Study 3, these differences all but disappeared.

Cross-task correlations between PAB and APTS tests

There were too few subjects in these studies to conduct a factor analysis, but cross-task correlations provide information about the degree to which individual tests from the two batteries share common variance and thus can imply which tests might be considered redundant when constructing a test battery aimed at representing unique factors. Table 6 displays a summary of the intercorrelation matrix for the most stable trial scores between PAB and APTS number correct measures after the Spearman-Brown and correction-for- attenuation formulas have been applied. On the average,

half the correlations between both batteries are $r < 0.25$, implying relative independence of tests. Table 7 reports comparable cross-task correlations between 15 PAB and 9 APTS using the response latency measure. Percent correct has been dropped as a metric. In general, the within-battery correlation (not shown) comparing number correct correlations are positive but generally lower than $r = 0.25$, and this is approximately the same level of correlations between tests of the two batteries. Therefore, any test could be combined with almost any other test in order to form an ad hoc battery.

Table 5.

Comparison of estimated test reliabilities across three studies of PAB and APTS

Test	Study 1 NC*	Study 1 PC	Study 1 RL	Study 2 NC*	Study 2 PC	Study 2 RL	Study 3 NC*	Study 3 RL
PAB								
Code Substitution				.67	.33	--		
Matrix Rotation	.67	.55	.51	.52	.69	.46	.89	.90
Recall	.72	--	.77					
Math Processing	.64	.51	.66				.83	.85
Item Order	.34	.45	.63					
Memory Search	.43	--	.61				.82	.78
Pattern Comp (Succ)	.38	--	.68					
Reasoning (Gram)				.72	.46	.68		
Reaction (4)						--		
Symbolic Reasoning							.85	.77
Vertical Addition							.72	.75
Time Wall								.82
Pattern Comp (Simult)							.62	.76
Manikin							.78	.68
Visual Scanning				.70	.45	.73		
APTS								
Tapping (Pref)	.99			.99			.98	
Pattern Comp (Simult)	.81	.49	.72	.82	.69	.74	.72	.71
Manikin	.97	.99	.97	.82	.95	.84	.86	.86
Reasoning (Gram)	.95	.86	.87	.71	.56	.83	.88	.88
Reaction Time (4)			.90			.70		.77
Reaction Time (2)			.84			.73		
Code Substitution	.71	.50	.80	.81	.28	.85	.82	.86
Tapping (Nonpref)	.99			.99			.99	
TF Tapping	.96			.99			.99	

*Tapping tests measures as number of alternate hits.
**Did not stabilize.
Note: NC = Number Correct; PC = Percent Correct; RL = Response Latency

Table 6.

Summary of cross-task correlations between PAB and APTS tests with Spearman-Brown and attenuation formulas applied (number correct measure*)

	Study 1			Study 2			Study 3		
Test	≥ .50	.49-.26	≤ .25	≥ .50	.49-.26	≤ .25	≥ .50	.49-.26	≤ .25
PAB									
Code Substitution				0	7	10			
Matrix Rotation	3	6	8	3	4	10	0	5	7
Recall	3	3	11						
Math Processing	0	4	13				3	4	5
Item Order	3	5	9						
Memory Search	0	6	11				4	3	5
Pattern Comp (Succ)	4	9	4						
Reasoning (Gram)				7	3	7			
Reaction (4)				-	-	-			
Symbolic Reasoning							6	0	6
Vertical Addition							0	2	10
Time Wall							-	-	-
Pattern Comp (Simult)							3	5	4
Manikin							2	6	4
Visual Scanning				3	2	12			
Average PAB	2.2	5.5	9.3	3.3	4.0	9.8	2.6	3.6	5.9
Percent Overlap	13%	32%	55%	19%	24%	57%	21%	30%	49%
APTS									
Tapping (Pref)	0	4	14	2	7	8	0	0	15
Pattern Comp (Simult)	0	7	13	5	6	6	0	8	7
Manikin	2	4	14	6	5	6	9	4	2
Reasoning (Gram)	4	5	11	4	4	9	4	10	1
Reaction Time (4)	-	-	-	-	-	-	-	-	-
Code Substitution	5	9	6	5	9	3	3	3	9
Tapping (Nonpref)	1	7	10	2	5	10	0	5	10
TP Tapping	0	8	12	0	2	15	0	3	12
Average APTS**	3.0	6.3	11.4	3.4	5.4	8.1	2.3	4.7	8.0
Percent Overlap	11%	32%	59%	20%	32%	48%	15%	31%	53%

*Tapping tests measures as number of alternate hits.
**Did not stabilize.
Note: NC = Number Correct; PC = Percent Correct; RL = Response Latency

Table 7.

Cross-task correlations between PAB and APTS tests with Spearman-Brown and attenuation formulas applied (response latency measure*)

Test	Study 1 ≥ .50	Study 1 .49-.26	Study 1 ≤ .25	Study 2 ≥ .50	Study 2 .49-.26	Study 2 ≤ .25	Study 3 ≥ .50	Study 3 .49-.26	Study 3 ≤ .25
PAB									
Code Substitution				4	8	5			
Matrix Rotation	3	5	9	6	3	8	2	4	6
Recall	0	2	15						
Math Processing	0	3	14				3	3	6
Item Order	0	7	10						
Memory Search	0	8	9				3	4	5
Pattern Comp (Succ)	2	6	9						
Reasoning (Gram)				9	1	7			
Reaction (4)				9	5	3			
Symbolic Reasoning							4	3	5
Vertical Addition							0	1	11
Time Wall							0	2	10
Pattern Comp (Simult)							4	4	4
Manikin							2	7	3
Visual Scanning				3	3	11			
Average PAB	0.8	5.2	11.0	6.2	4.0	6.8	2.3	3.5	6.3
Percent Overlap	5%	31%	65%	36%	24%	40%	19%	29%	52%
APTS									
Tapping (Pref)	-	-	-	-	-	-	-	-	-
Pattern Comp (Simult)	8	4	8	5	1	10	2	6	7
Manikin	3	6	11	4	10	3	11	1	3
Reasoning (Gram)	3	5	12	5	2	10	5	8	2
Reaction Time (4)	2	8	10	8	4	5	0	2	13
Code Substitution	4	5	11	6	6	5	2	4	9
Tapping (Nonpref)	-	-	-	-	-	-	-	-	-
TF Tapping	-	-	-	-	-	-	-	-	-
Average APTS**	4.0	5.6	10.4	5.6	4.6	6.6	4.0	4.2	6.8
Percent Overlap	20%	28%	52%	33%	27%	39%	27%	28%	45%

*Tapping tests measures as number of alternate hits.
**Did not stabilize
Note: NC = Number Correct; PC = Percent Correct; RL = Response Latency

From the PAB battery, it appears that Code Substitution, Math Processing, Memory Search, and Vertical Addition exhibited the least overlap with APTS tests, considering the number correct measure (Table 6). For the APTS battery, any of the Tapping tests exhibited relative independence, although they correlated almost perfectly with each other. Simultaneous Pattern Comparison shared the least variance with PAB tests and Manikin evidenced the greatest overlap with PAB tests. The highest correlations tended to be between tasks from the two batteries which are essentially the same (e.g., Manikin and Grammatical Reasoning). It should be noted that, because of our greater experience with APTS tests, the plan of the studies was to use the APTS core battery repeatedly in the three studies. However, this also overemphasizes somewhat the overlap. In another study (Kennedy, Baltzley, Wilkes, & Dunlap, 1989) 15 additional APTS tests are available with the prospect of greater factorial diversity. Overall, across the three studies, PAB tests averaged 54% cross-task correlations of 0.25 or less with APTS, and APTS tests had an average of 42% similarly low cross-task correlations with PAB tests.

For the response latency measure (Table 7), again using the number of correlations of 0.25 or less as a criterion, PAB tests overlapped less with APTS tests (average = 53%) than APTS tests overlapped with PAB tests (average = 45%). From the PAB battery, tests which exhibited the least overlap with APTS tests were Recall, Math Processing, Item Order, Vertical Addition, and Time Wall. None of the APTS tests exhibited comparable correlational independence from PAB tests across the three studies. There is no apparent reason why cross-task correlations were relatively high in Study 2, except for the fact that perhaps the trials (9 and 10) upon which cross-task correlations were based occurred just prior to a one-week no-test interval, so possibly unique performance predominated.

Conclusions

Throughout this experimental program to select the "best" tests for an optimal computerized test battery for assessment of environmental effects on skilled behavior and higher level tasks, we have stressed the need for repeated-measures experiments to properly evaluate test stability, reliability, and factorial purity. The three repeated-measures studies conducted for this program evaluated 15 tests from the PAB in comparison with nine marker tests from the APTS (three of which are essentially redundant -- Tapping tests). So far as we know, the PAB tests had not previously been implemented or evaluated on portable microcomputers.

In general, we found that improved experimental design and administrative procedures led to greater comparability between the two test batteries as lessons were learned from early studies. For example, later studies counterbalanced or equated test presentation, provided a computerized "Smart" system to warn experimenters when subjects were responding poorly in practice, and developed an objective approach to the determination of the trial at which differential stability was achieved.

In spite of some experimental and administrative differences across studies, it is possible to reach conclusions regarding which tests and scores would be recommended to include in an optimal battery.

The literature (Dunlap, Kennedy, Harbeson, & Fowlkes, in press; Carter, Krause, & Harbeson, 1986) suggests that percent correct scores are likely to be less reliable and we found this to be so. While we advocate the use of percent correct as a check on the consistency of the strategy employed by the subjects, we believe percent correct should generally not be used in statistical analysis of treatment effects unless due care is taken of their generally lower reliabilities and the effect such would have on power. Therefore, disregarding percent correct response measures and using only the more stable and reliable number correct and response latency measures, we find that all tests of the APTS exceeded our 0.707 criterion for acceptable retest reliability. They are Simultaneous Pattern Comparison, Manikin, Grammatical Reasoning, Four-Choice Reaction Time, Two-Choice Reaction Time, Code Substitution, and all Tapping tests. Seven PAB tests exceeded the criterion for both response measures. They are Recall, Mathematical Processing, Grammatical Reasoning, Symbolic Reasoning, Visual Scanning, Vertical Addition, and Time Wall.

It is noteworthy that none of the five PAB tests which assessed the same domains as the APTS tests (e.g., Code Substitution, Manikin, etc.) reached comparably high levels of reliability. Although empirical checks have been made successfully for microcomputer tests with the Spearman-Brown (Kennedy, Carter, & Bittner, 1980), we do not really know if "time" of performance can be substituted directly for items attempted in the Spearman-Brown formula. If stabilized reliability is the prime criterion for test selection we believe PAB and APTS tests are comparable, but if amount of time invested in practice and performance is a consideration, and if one wishes to achieve the most rapid stabilization, then the APTS tests clearly have an advantage since they stabilize in approximately half as many minutes. Although lengthening a test such as the APTS will usually increase reliability, no empirical data exist to support this assumption in these particular studies, and fatigue during the longer (e.g., three minute) test of the PAB series might actually somewhat lower the reliability. Conceivably, the longer test might also prove more sensitive through this same mechanism. Many studies of these relations still need to be performed.

It is also worth noting that the PAB tests administered in two studies (Matrix Rotation, Mathematical Processing, and Memory Search) exhibited considerably greater reliability and exceeded the 0.707 criterion in the second testing which took place in Study 3. There were several additions in this study. These were procedural improvements (e.g., the Smart system -- objective stability determination) and more diverse subjects (handedness and gender) were used. These PAB and APTS tests stabilize rapidly, generally within three test trials or under nine minutes of testing time.

After correction for attenuation due to unreliability, an analysis of cross-task correlations revealed that PAB tests generally showed less correlational overlap with APTS tests than APTS did with PAB, suggesting that PAB tests are factorially more unique. But also their reliability correlations are lower and there are almost twice as many tests. PAB tests were studied as were APTS. Moreover, the same APTS marker tests were repeatedly employed in all the studies, and all but Tapping appeared in PAB in a different version. The other dozen APTS tests were not included (Kennedy, Baltzley, Wilkes, & Kuntz, 1989) in these studies and may tap other cognitive constructs. It would appear that the "core" battery of APTS tests is just that and it would be possible to add other tests (e.g., Recall and Math Processing) from the PAB in order to build a battery in a larger sample for factor analytic studies.

Those tests which exhibited the least overlap with other tests were APTS' Simultaneous Pattern Comparison and any of the Tapping series, and PAB's Mathematical Processing and Vertical Addition. In another study (Kennedy, Jones, Baltzley, & Turnage, 1988), 10 tests from PAB and APTS were factor analyzed and the authors recommended a core battery consisting of Recall (PAB), Matrix Rotation (PAB), Grammatical Reasoning (APTS), Mathematical Processing (PAB), Pattern Comparison (APTS), and the Tapping (APTS) or Reaction Time (APTS) tests. This battery was recommended based on three factors which emerged consistently across three test administrations: a verbal/spatial factor (identified by Recall, Grammatical Reasoning, and Matrix Rotation tests), a perceptual/numerical factor (identified by Math Processing, Pattern Comparison, and Code Substitution tests), and a motor speed factor (identified by Reaction Time and Tapping tests). Further factor analytic studies with larger test batteries and larger subject pools are likely to yield more factors, providing a rapidly administered final battery which is optimal in the sense of stability, reliability, and factorial richness. Some of this work has been completed and analyses proceed (Lane & Kennedy, 1988).

Combining what we have learned from the three studies summarized herein, as well as well as the Lane and Kennedy (1988) factor analytic study, we would tentatively suggest that the shortest optimal five-test battery taking 8-10 minutes should consist of Nonpreferred Hand Tapping and the APTS 4-Choice Reaction, APTS Code Substitution, Grammatical Reasoning, and APTS Pattern Comparison (or PAB Reaction Time, Code Substitution, Grammatical Reasoning, and Simultaneous Pattern Comparison). For 6-,7-,8-,9-,10-,11- and 12-test batteries, each taking an additional three minutes' administration time, one would add APTS, Manikin, Two-Finger Tapping, PAB Math Processing, PAB Simultaneous Pattern Comparison, PAB Spatial Processing (or PAB Successive Pattern Comparison), PAB Symbolic Reasoning, and APTS 2-Choice Reaction Time (or PAB Reaction Time), respectively. These tests exhibit high reliability, early stability, and factorial richness.

Although cognitive theory has served to surface tests for various batteries, none has been as comprehensively developed as were the tests of the French-Ekstrom-Price series (1963) which combined cognitive theory and classical test theory. However, that battery, while a formidable undertaking, had primary and secondary selection as is applied purpose and so only one or two forms were available for each construct. Because repeated-measures is an ordinary consequence of environmental stress studies, we believe it is correct to say that nc microcomputer-based battery of tests has had both cognitive theory and test theory guide its development. Hunt (1985) has bemoaned this situation on philosophical grounds and has suggested that cognitive theory may need tc guide develcpment. We used to think that test theory should guide development but are prepared to soften that position after 10 years in that mode. We believe the core battery from this paper plus the factor analystic study should be subjected to a theoretical decomposition and future tests proposed from theory.

In summary, we believe that these three studies have contributed useful information regarding the psychometric soundness of all the tests which were evaluated for possible inclusion in a portable, computerized repeated-measures battery of tests for the assessment of environmental effects on skilled behavior and higher level tasks.

References

Allen, M. J., and Yen, W. M. 1979. Introduction to measurement theory. Monterey, CA: Brooks-Cole Publishing.

Alvares, K. M., and Hulin, C. L. 1972. Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. Human factors. 14:295-308.

American Psychological Association. 1982. Ethical principles in the conduct of research with human participants. Washington, DC: American Psychological Association.

Anderson, E. B. 1985. Estimating latent correlations between repeated testing. Psychometrika. 50:3-16.

Bachrach, A. J. 1975. Underwater performance. In P. B. Bennett and D. H. Elliot, eds. The physiology and medicine of diving and compressed air work, 2nd ed., pp. 274-284. London: Bailliere Tindall.

Baker, E. L., Letz, R. E., Fidler, A. T., Shalat, S., Pantamura, O., and Lyndon, M. 1985. A computer-based neurobehavioral evaluation system for occupational and environmental epidemiology: Methodology and validation studies. Neurobehavioral toxicology and teratology. 7:369-377.

Bandaret, L. E., MacDougall, D. M., Roberts, D. E., Tappan, D., Jacey, M., and Gray, P. 1984. Effects of dehydration or cold exposure and restricted fluid intake upon cognitive performance. Proceedings of the National Academy of Sciences (Committee on Military Nutrition Research Workshop), "Workshop on knowledge needed for the development of predictive models of military performance decrements resulting from inadequate nutrition." Washington, DC: National Academy of Sciences.

Barrett, G. V., Alexander, R. A., and Forbes, J. B. 1977. Analysis of performance measures. JSAS catalogue of selected documents in psychology. 7:1623.

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., and Krause, M. 1983. Repeated measurements of human performance. New Orleans, LA: Naval Biodynamics Laboratory, NBDL-83R006

Bittner, A. C., Jr., Smith, M. G., Kennedy, R. S., Staley, C. F., and Harbeson, M. 1985. Automated portable test (APT) system: Overview and prospects. Behavior research methods, instruments, & computers. 17:217-221.

Carter, R. C., Kennedy, R. S., and Bittner, A. C., Jr. 1980. Selection of performance evaluation tests for environmental research. Proceedings of the 24th annual meeting of the Human Factors Society, pp. 320-324. Santa Monica, CA: Human Factors Society. Also, New Orleans, LA: Naval Biodynamics Laboratory, July 1981, pp. 1-7, NBDL-81-R0088. (DTIC No. AD A111296)

Carter, R. C., Kennedy, R. S., and Bittner, A. C., Jr. 1981. Grammatical reasoning: A stable performance yardstick. Human factors, 23:587-591.

Carter, R. C., Krause, M., and Harbeson, M. M. 1986. Beware the reliability of slope scores for individuals. Human factors, 28:673-683.

Deroshia, C. 1989. The effect of exercise and training upon performance and mood during antiorthostatic bed rest. Moffett Field, CA: Ames Research Center. In press.

Dunlap, W. P., Jones, M. B., and Bittner, A. C., Jr. 1983. Average correlations vs. correlated averages. Bulletin of the psychonomic society, 21:213-216.

Dunlap, W. P., Kennedy, R. S., Harbeson, M. M., and Fowlkes, J. E. (1989). Difficulties with individual difference measures upon recent cognitive paradigms. Applied psychological measurement journal. In press.

Ellis, H. D. 1982. The effects of cold on the performance of serial choice., reaction time, and various discrete tasks. Human factors, 24(5):589-598.

Englund, C. E., Reeves, D. L., Shingledecker, C. A., Thorne, D. R., Wilson, K. P., and Hegge, F. W. 1986. Unified Tri-Service cognitive performance assessment battery (UTC-PAB). Fort Dietrick, MD: U.S. Army Research and Development Command. Report No. 86-1.

Englund, C. E., Ryman, D. H., Naitoh, P., and Hodgdon, J. A. 1985. Cognitive performance during successive sustained physical work episodes. Behavior research, methods, instruments, & computers, 17(1):75-85.

Essex Corporation. 1981, July. Isoperformance technical memorandum No. 5: A demonstration program. Orlando, FL: Essex Corporation.

Fleishman, E. A., and Hempel, W. E., Jr. 1955. The relation between abilities and improvement with practice in a visual discrimination reaction task. Journal of experimental psychology, 49:301-312.

Fowler, B., Paul, M., Porlier, G., Elcombe, D. D., and Taylor, M. 1985. A reevaluation of the minimum altitude at which hypoxic performance decrements can be detected. Ergonomics. 28:781-791.

French, J.W., Ekstrom, R.B., and Price, L.A. 1963. Manual for kit of reference tests for cognitive factors. Research Contract NONR-2214-00. Princeton, NJ: Educational Testing Service.

Guilford, J. P. 1954. Psychometric methods (2nd ed.). New York: McGraw-Hill, 400-402.

Gulliksen, H. 1950. Theory of mental tests. New York: Wiley.

Guillion, C. M., and Eckerman, D. A. 1986. Field testing for neuro-behavioral toxicity: Methods and methodological issues. In Z. Annau (Ed.), Behavioral toxicology. Baltimore, MD: Johns Hopkins Press.

Hamilton, B., Simmons, R., and Kimball, K. 1982. Biological effects of chemical defense ensemble-imposed heat stress of Army aviators. Fort Rucker, AL: U.S. Army Aeronautical Research Laboratory. USAARL Report No. 83-F.

Hettinger, L. J., Kennedy, R. S., and McCauley, M. E. 1988. Motion and human performance. In G. H. Crampton, ed. Motion and space sickness. Boca Raton, FL: CRC Press.

Hunt, E. 1985. Science, technology, and intelligence. Buros Institute Annual Conference, Lincoln, NE. Report No. 9, Grant No. N00014-84-K-553.

Johnson, J. H., and Kennedy, R. S. 1985. Literature review and critique of methods to assess human performance in dynamic vehicle/operator settings, Task 1. Fort Dietrick, Frederick, MD: U.S. Army Medical Research Acquisition Activity. Contract No. DAMD-85-C-5095.

Jones, M. B. 1970. Rate and terminal processes in skill acquisition. American journal of psychology. 83(2):222-236.

Jones, M. B. 1979. Stabilization and task definition in a performance test battery. New Orleans, LA: U.S. Naval Aerospace Medical Research Laboratory. Final Report, Contract N00203-79-N-5089.

Jones, M. B. 1980. Sequential precession and diminishing returns in the acquisition of a motor skill. Journal of motor behavior, 12:69-73.

Jones, M.B., Dunlap, W.P., and Bilodeau, I.M. 1984. Factors appearing late in practice. Organizational behavior and human performance. 33:153-173.

Keatinge, W. R. 1969. Survival in cold water. Oxford, England: Blackwell Scientific Publications.

Kennedy, R. S., Baltzley, D. R., Wilkes, R. L., and Kuntz, L. A. 1989. A menu of self-administered microcomputer-based neurotoxicology tests. Behavior research methods, instruments, & computers. Submitted for publication.

Kennedy, R. S., and Bittner, A. C., Jr. 1977. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). Personnel performance measurement symposium. L. T. Pope and D. Meister (Eds.) Productivity enhancement: Personnel performance assessment in Navy systems. San Diego, CA: Navy Personnel Research and Development Center. (DTIC No. AD A056074)

Kennedy, R. S., Carter, R. C., and Bittner, A. C., Jr. 1980. A catalogue of performance evaluation tests for environmental research. Proceedings of the Human Factors Society, pp. 334-348. Los Angeles, CA.

Kennedy, R. S., and Frank, L. H. 1986. A review of motion sickness with special reference to simulator sickness. Presented at the 65th Annual Meeting of the Transportation Research Board, Washington, DC.

Kennedy, R. S., Jones, M. B., Baltzley, D. R., and Turnage, J. J. 1988. Factor and regression analysis of a microcomputer-based cognitive test battery. Orlando, FL: Essex Corporation.

Kennedy, R. S., Wilkes, R. L., and Baltzley, D. R. 1989. Microcomputer-based mental acuity tests indexed to alcohol dosage. Orlando, FL: Essex Corporation. In press.

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., and Kuntz, L. A. 1987. Microbased repeated-measures performance testing and general intelligence. Presented at the 29th annual conference of the Military Testing Association. Ottawa, Ontario, Canada.

Kennedy, R.S., Wilkes, R.L., and Kuntz, L.A. 1986. Sensitivity of a notebook-sized portable automated performance test system. Presented at the annual Behavioral Toxicology Society meeting. Atlanta, GA.

Kennedy, R. S., Wilkes, R. L., Lane, N. E., and Homick, J. L. 1985. Preliminary evaluation of microbased repeated-measures testing system. Orlando, FL: Essex Corporation. Report EOTR-85-1 for NASA Johnson Space Center.

Kohl, R. L., Calkins, M. A., and Mandell, A. J. 1986. Arousal and stability: The effects of five new sympathomimetic drugs suggest a new principle for the prevention of space motion sickness. Aviation, space, & environmental medicine, 57:137-143.

Kryter, K.D. 1970. The effects of noise on man. New York: Academic Press.

Lane, N.E. 1986. Issues in performance measurement for military aviation with applications to air combat maneuvering. Orlando, FL: Naval Training Systems Center. NTSC TR-86-008.

Lane, N. E., and Kennedy, R. S. (Eds.). 1988). Users manual for the U.S. Army Aeromedical Research Laboratory Portable Performance Assessment Battery. Fort Rucker, AL: U.S. Army Aeromedical Laboratory. USAARLPPAB Tech. Report No. EOTR 88-5.

Logie, R. H., and Baddeley, A. D. 1985. Cognitive performance during simulated deep-sea diving. Ergonomics, 28:731-746.

McCauley, M.E., Kennedy, R.S., and Bittner, A.C., Jr. 1980. Development of Performance evaluation tests for environmental research (PETER): Time estimation. Perceptual and motor skills, 51:655-665.

Mitchell, G., Knox, F., Edwards, R., Schrimscher, R., Siering, G., Stone, L., and Taylor, P. 1985. Microclimate cooling and the aircrew chemical defense ensemble. Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

Mohs, C., Tinklenburg, J. R., Roth, W. T., and Kopell, B. S. 1970. Sensitivity of some human cognitive functions to the effects of methamphetamine and secobarbital. Drug and alcohol dependence, 53:145-150.

Naitoh, P. (1982). Chronobiologic approach for optimizing human performance. In F. J. Brown and R. C. Gaeber (Eds.), Rhythmic aspects of behavior, pp. 41-102. Hillsdale, NJ: Erlbaum.

NEC Home Electronics (USA), Inc. 1983. NEC PC-8201A users guide. Tokyo: Nippon Electric Co., Ltd.

Nicogossian, A. E., and Parker, J. F. 1982. Space physiology and medicine. Houston, TX: NASA, Scientific and Technical Information Branch.

Perez, W. A., Masline, P. J., Ramsey, E. G., and Urban, K. E. 1987. Unified Tri-Service cognitive performance assessment battery: Review and methodology. Wright-Patterson AFB, OH: Armstrong Aerospace Medical Research Laboratory. Report No. AAMRL-TR-87-007.

Reeves, D. L., and Thorne, D. R. 1988. Development and application of the unified tri-service cognitive assessment battery within naval aviation. Presented at the 59th annual scientific meeting of the Aerospace Medical Association. New Orleans, LA.

Rognum, T. O., Vartdal, F., Rodahl, K., Opstad, P. K., Knudson-Bass, O., Kindt, E., and Withey, W. R. 1986. Physical and mental performance of soldiers on high- and low-energy diets during sleep deprivation. Ergonomics, 29:859-867.

Sanders, A. F., Haywood, R. C., Schroiff, H. W., and Wauschkuhn, C. H. 1986). Standardization of performance tests: A proposal for further steps. U.S. Air Force Office of Scientific Research. Report No. EOTR-TR-86,08. (NTIS No. AD A172682)

Seales, D. M., Kennedy, R. S., and Bittner, A. C., Jr. 1980. Development of performance evaluation tests for environmental research (PETER): Arithmetic computation. Perceptual and motor skills. 51:1023-1031.

Shingledecker, C. A. 1984. A task battery for applied human performance assessment research. Dayton, OH: Air Force Aerospace Medical Research Laboratory. AFAMRL-TR-84.

Shoenberger, R. W. 1981. Application of psychophysical methods to judgements of whole-body vibration intensity. Presented at the International workshop on Research Methods in Human Motion and Vibration Studies. New Orleans, LA.

Smith, A., and Miles, C. 1986. The effects of lunch on cognitive vigilance tasks. Ergonomics:29:1251-1261.

Spearman, C. 1904. The proof and measurement of association between two things. American journal of psychology. 15:72-101.

Tabler, R. E., Turnage, J. J., and Kennedy, R. S. 1987. Repeated-measures analyses of psychomotor tests from the APTS battery and selected PAB tests: Stability, reliability and cross-task correlations. Study 2. Orlando, FL: Essex Corporation.

Thorndike, R. L., and Hagen, E. P. 1977. Measurement and evaluation in psychology and education, 4th ed. New York: Wiley.

Thorne, D. 1982. Documentation: The Walter Reed performance assessment battery. Washington, DC: Walter Reed Army Institute of Research. U.published.

Thorne, D., Genser, S. G., Sing, H. C., and Hegge, F. W. 1985. The Walter Reed performance assessment battery. Neurobehavioral toxicology & teratology. 7:415-418.

Turnage, J. J., Kennedy, R. S., and Osteen, M. K. 1987. Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations. Orlando, FL: Essex Corporation.

Turnage, J. J., Kennedy, R. S., Osteen, M. K., and Tabler, R. E. 1988. Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations: Study 3. Orlando, FL: Essex Corporation.

West, V., and Parker, J. F. 1975. A review of recent literature: Measurement and prediction of operational fatigue. Arlington, VA: Office of Naval Research. ONR Report No. 201-067. (DTIC No. AD A008405)

Wilkes. R. L., Kennedy, R. S., and Kuntz, L. A. 1987. A comparison of NEC and Zenith microcomputer administrations of a battery of human performance tests. Orlando, FL: Essex Corporation. Unpublished.

Winer, B. J. (1971). Statistical principles in experimental design, 2nd ed. New York: McGraw Hill.

Woodward, D. P., and Nelson, P. D. 1974. A user oriented review of the literature on the effects of sleep loss, workrest schedules, and recovery on performance. Washington, DC: Office of Naval Research. ONR Report No. ACR 206. (DTIC No. AD A009778)